Proceedings of the Society for Computation in Linguistics

Volume 3 Article 46

2020

Information-theoretic Characterization of the Sub-regular Hierarchy

Huteng Dai
Rutgers University, huteng.dai@rutgers.edu

Richard Futrell
University of California, Irvine, rfutrell@uci.edu

Information-theoretic characterization of the Sub-regular Hierarchy

Huteng Dai
Rutgers University
huteng.dai@rutgers.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Our goal is to link two different formal notions of complexity: the complexity classes defined by **Formal Language Theory** (FLT)—in particular, the Sub-regular Hierarchy (Rogers et al., 2013; Lai, 2015; Heinz, 2018)—and **Statistical Complexity Theory** (Feldman and Crutchfield, 1998; Crutchfield and Marzen, 2015). The motivation for exploring this connection is that factors involving memory resources have been hypothesized to explain why phonological processes seem to inhabit the Sub-regular Hierarchy, and Statistical Complexity Theory gives an information-theoretic characterization of memory use. It is currently not known whether statistical complexity and FLT define equivalent complexity classes, or whether statistical complexity cross-cuts the usual FLT hierarchies. Our work begins to bridge the gap between FLT and Information Theory by presenting characterizations of certain Sub-regular languages in terms of statistical complexity.

Statistical complexity theory. Statistical complexity theory deals with stochastic processes: probabilistic models of infinitely-long sequences. For a process X generating sequences of symbols indexed as $\ldots X_{t-2}, X_{t-1}, X_t, X_{t+1}, \ldots$, we define the notation $\overleftarrow{X}$ ("**the past**") to mean $\ldots X_{t-2}, X_{t-1}$, and $\overrightarrow{X}$ ("**the future**") to mean $X_t, X_{t+1}, \ldots$.

The **statistical complexity** of a stochastic process is the minimal amount of information about the past required to faithfully reproduce the future. Suppose that we want to simulate a stochastic process by generating each symbol based on some memory representation M of the past, and that we want to find a memory representation M that simulates the process as well as possible while having minimal information content, measured in bits. This quantity of minimal information is called statistical complexity. Formally, the statistical complexity S of a process X is the minimum entropy of a memory representation M that perfectly simulates the process:

$$S \equiv \min_{M: D_{KL}[\overrightarrow{X}|\overleftarrow{X}||\overrightarrow{X}|M]=0} H[M], \qquad (1)$$

where $H[M]$ is the entropy of the random variable M:

$$H[M] \equiv -\sum_x p_M(x) \log p_M(x), \qquad (2)$$

and $D_{KL}[\cdot||\cdot]$ is conditional Kullback-Leibler divergence (see Cover and Thomas, 2006), which is zero for identical conditional distributions. Therefore, Eq. 1 indicates the minimum entropy of any memory representation M subject to the constraint that M must allow us to generate a distribution over future sequences $\overrightarrow{X}$ which is identical to the distribution we would have generated given the past $\overleftarrow{X}$.

Further insight comes from considering the different factors that contribute to statistical complexity. Using information-theoretic identities, we break the statistical complexity into two terms:

$$\begin{aligned} S = H[M] &= I[M : \overrightarrow{X}] + H[M|\overrightarrow{X}] \\ &= \underbrace{I[\overleftarrow{X} : \overrightarrow{X}]}_{E} + \underbrace{H[M|\overrightarrow{X}]}_{C}, \end{aligned}$$

where $I[\cdot : \cdot]$ is mutual information, the amount of information in one random variable about another. The term E is called **excess entropy** and quantifies the amount of information in the past which is useful for predicting the future. The term C is called **crypticity** and quantifies the amount of information stored in M which does not end up being useful for predicting the future.

These quantities are easily understood in terms of memory resources used for incremental language production and comprehension. Statistical complexity measures memory load or storage

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 445-448.
New Orleans, Louisiana, January 2-5, 2020

cost; it can be finite even for non-finite-state processes, as long as the sum in Eq. 2 converges. Excess entropy measures integration cost: it says how many bits of information from the past are used when processing the future. Crypticity is the difference between statistical complexity and excess entropy, and measures the amount of information stored in the minimal memory representation M which does not ultimately end up being used to predict the future.

In order to study memory efficiency, we use these quantities to define an efficiency metric, the **E/S ratio**, which is excess entropy divided by statistical complexity. The E/S ratio tells the proportion of bits stored in memory which end up being useful for predicting the future.

Preliminaries. We study Sub-regular languages defined using Probabilistic Deterministic Finite-state Automata (PDFAs). A PDFA is characterized by a set of internal states $\mathcal{Q}$, an alphabet Σ, an **emission distribution** O of symbols $\in \Sigma$ conditional on a state $\in \mathcal{Q}$, a **transition function** $T : \mathcal{Q} \times \Sigma \to \mathcal{Q}$ defining which state the machine transitions into after emitting a symbol, and distinguished initial and final states. In a PDFA, the transition function T is deterministic; in a general Probabilistic Finite-state Automaton (PFA), it may be stochastic, in which case we have a **transition distribution** rather than a transition function. Our indexing convention is: at time t, the PDFA is in state q_t; it generates symbol x_t before transitioning into the next state q_{t+1}. The time indexing convention is shown in Figure 1.

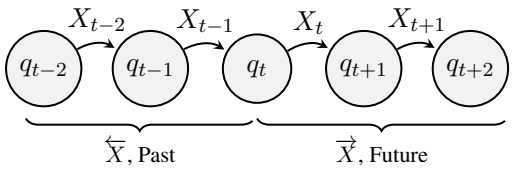

Figure 1: Time-indexing conventions for a finite-state machine.

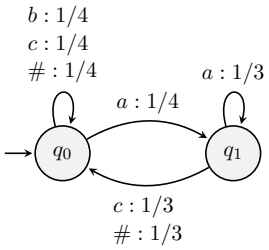

Figure 2: SL_2 PDFA of $\neg ab$, $\Sigma = \{a, b, c\}$

We use the following construction to generate a stationary ergodic stochastic process from a PDFA: whenever the PDFA emits an end-of-word symbol #, it always transitions back into the initial state. The resulting infinite stream of symbols is amenable to analysis using statistical complexity theory. In the literature on statistical complexity, a PDFA of this form is called a **unifilar Hidden Markov Model** (Travers and Crutchfield, 2011, unifilar HMM).

Below, we describe how to calculate S, E, and C from the **minimal trimmed** PDFA (Heinz and Rogers, 2010) for Strictly k-Local (SL_k) languages.

Statistical complexity. For a unifilar HMM, the statistical complexity reduces to the entropy of the stationary distribution over internal states (Travers and Crutchfield, 2011). To get the stationary distribution over internal states Q, we first construct a **state transition matrix**: a stochastic matrix whose entries represent the probability of going into state q_{t+1} after being in state q_t. For a general PFA, the entries of this matrix are given by marginalizing over the emission distribution O:

$$p(q_{t+1}|q_t) = \sum_{x_t \in \Sigma} p_O(x_t|q_t) p_T(q_{t+1}|x_t, q_t),$$

where p_T is the probability of transitioning into state q_{t+1} after generating symbol x_t from state q_t. In a PDFA, this probability is given by the deterministic transition function T, so the transition probability p_T reduces to a Kronecker delta function:

$$p_T(q_{t+1}|x_t, q_t) = \delta_{q_{t+1}=T(x_t,q_t)}.$$

Finally, the stationary distribution over states Q is given by the left eigenvector of the state transition matrix associated with eigenvalue 1.

In general, the statistical complexity of a process depends on the minimal number of states required to represent the process as a PDFA. For an SL_k language, statistical complexity is upper bounded as $S \leq (k-1) \log |\Sigma|$.

Excess entropy. For SL_k languages,

$$E = I[X_{t-k+1}, \ldots, X_{t-1} : X_t, \ldots, X_{t+k-2}].$$

In the case of SL_2 languages, we compute E by constructing a **symbol transition matrix**, a stochastic matrix whose entries represent

$p(x_{t+1}|x_t)$, marginalizing over q_t and q_{t+1}. We also need the stationary distribution over symbols, derived from the symbol transition matrix by the same procedure as above.

Crypticity. Crypticity $C = S - E$. In general, crypticity is bounded above by the uncertainty about the emitting state given a symbol:

$$C \leq H[Q_t|X_t],$$

with equality iff X is an SL_2 language.

Sub-regular Hierarchy. We consider two relational structures, namely the successor ($+1$) and precedence ($<$) relations. Languages with successor relation keep track of k-long **sub-strings** of the input, such as $\{aa, ab, ac, ba, \ldots\}$ in an SL_2 language. On the other hand, languages with precedence relation keep track of k-long **subsequences**, such as $\{a \ldots a, a \ldots b, \ldots\}$ in an SP_2 language. Different sub-regular languages correspond to distinct PDFAs. For each relational structure, languages with the higher logical power are considered to be more expressive. For example, SL languages are a subset of locally testable (LT) languages. The subset relations are indicated by lines connecting higher and lower regions in Figure 3.

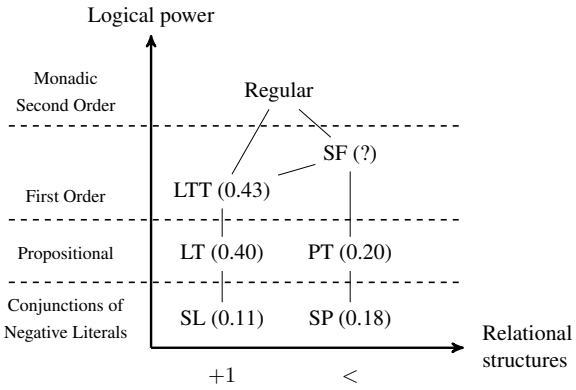

Figure 3: Sub-regular Hierarchy, with E/S ratios calculated from the examples in the text.

Table 1 shows calculated statistical complexity, excess entropy, and crypticity for the minimal trimmed PDFAs of example languages in the Sub-regular Hierarchy, including Strictly Local (SL), Locally Testable (LT), Locally Threshold Testable (LTT), Strictly Piecewise (SP), Piecewise Testable (PT).

The information quantities align with the hypothesis in FLT literature: the languages which are more expressive have higher memory storage requirements. E/S ratios characterize the subset relation in the Sub-regular Hierarchy, for both successor and precedence relations: the higher regions in the hierarchy have higher amount of E/S ratio, as illustrated in Figure 3.

	$\mathbf{SL}_2$	$\mathbf{LT}_2$	$\mathbf{LTT}_2$	$\mathbf{SP}_2$	$\mathbf{PT}_2$
Statistical complexity	0.97	1.53	1.94	0.99	1.53
Excess entropy	0.09	$\geq$0.61	$\geq$0.83	$\geq$0.18	$\geq$0.30
Crypticity	0.75	$\leq$0.91	$\leq$1.10	$\leq$0.80	$\leq$1.22
E/S ratio	0.11	$\geq$0.40	$\geq$0.43	$\geq$0.18	$\geq$0.20

Table 1: Information quantities for PDFAs shown in figures. SL_2 = Figure 2; LT_2 = Figure 4; LTT_2 = Figure 5, SP_2 = Figure 6; PT_2 = Figure 7. Quantities marked with $\leq$ or $\geq$ are bounds based on Markov approximations.

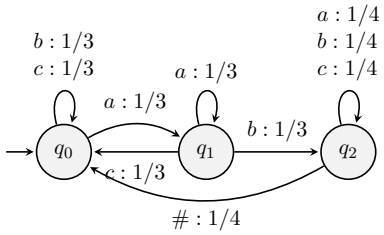

Figure 4: LT_2 PDFA of Some-ab, $\Sigma = \{a, b, c\}$

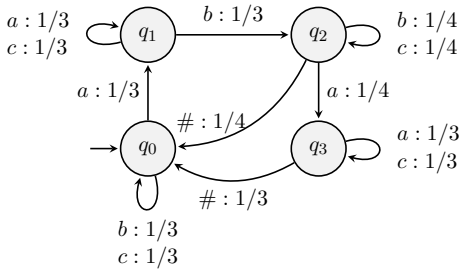

Figure 5: LTT_2 PDFA of One-ab, $\Sigma = \{a, b, c\}$

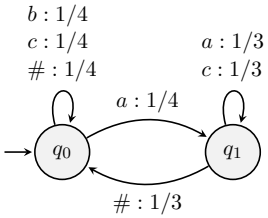

Figure 6: SP_2 PDFA of $\neg a \ldots b$, $\Sigma = \{a, b, c\}$

The information-theoretic characterization illuminates the comparison across relational structures. For example, SL and SP languages correspond to different types of phonotactics: SL

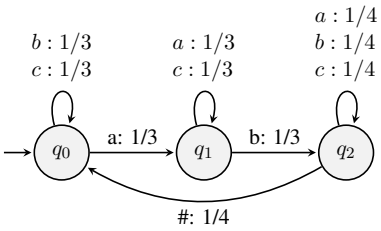

Figure 7: PT_2 PDFA of Some-$a \dots b$, $\Sigma = \{a, b, c\}$

only describes local phonotactics, while SP corresponds to patterns of long-distance agreement. In the examples we have examined, SL and SP have similar information quantities when they share the same k-factor. We conjecture that SL_k and SP_k languages have similar memory efficiency because they are both described by Conjunction of Negative Literals (McNaughton and Papert, 1971, CNL; the combination of $\neg$ and $\wedge$).

Conclusion. We have investigated whether there is a coherent relationship between complexity metrics calculated using Statistical Complexity Theory on one hand, and the Sub-regular hierarchy of languages on the other hand. Our preliminary results, based on example languages representing a number of Sub-regular classes, suggest that increasing logical power corresponds to increasing information-theoretic memory storage requirements. Our current study is limited in that we have only calculated complexity metrics for selected examples of each language class. Future work will work to establish general formal relationships between language classes and statistical complexity.

Regardless of whether statistical complexity turns out to map cleanly onto FLT hierarchies, we believe it provides a promising framework for characterizing bounds on complexity of human languages and phonotactics in particular. The theory of statistical complexity provides a clear way to quantify and reason about memory storage cost and memory integration cost in a highly general information-theoretic setting. Therefore it is entirely reasonable to expect that there may be bounds on the complexity of linguistic subsystems, defined using the language of statistical complexity.

In this connection, we note that statistical complexity depends on a number of factors that are not usually relevant in FLT, such as the transition probabilities and number of states in a PDFA. Although these factors are not relevant in FLT, they may nonetheless be relevant for characterizing constraints on the phonology and phonotactics of human languages. By characterizing complexity using Statistical Complexity Theory, we can take these factors into account in a principled way.

Acknowledgement. We thank Jim Crutchfield, Jeff Heinz, Adam Jardine, and anonymous reviewers for their comments and insights.

References

Thomas M. Cover and J.A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

James P Crutchfield and Sarah Marzen. 2015. Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning. *Physical Review E*, 91(5):050106.

David P. Feldman and James P. Crutchfield. 1998. Measures of statistical complexity: Why? *Physics Letters A*, 238(4-5):244–252.

Jeffrey Heinz. 2018. The computational nature of phonological generalizations. *Phonological Typology, Phonetics and Phonology*, pages 126–195.

Jeffrey Heinz and James Rogers. 2010. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 886–896, Uppsala, Sweden. Association for Computational Linguistics.

Regine Lai. 2015. Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, 46(3):425–451.

Robert McNaughton and Seymour A Papert. 1971. *Counter-Free Automata (MIT research monograph no. 65)*. The MIT Press.

James Rogers, Jeffrey Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert, and Sean Wibel. 2013. Cognitive and sub-regular complexity. In *Formal grammar*, pages 90–108. Springer.

Nicholas F. Travers and James P. Crutchfield. 2011. Asymptotic synchronization for finite-state sources. *Journal of Statistical Physics*, 145(5):1202–1223.